Approximate All–Pairs Suffix/Prefix Overlaps

Niko Välimäki

Department of Computer Science
University of Helsinki
nvalimak@cs.helsinki.fi

Join work with Susana Ladra and Veli Mäkinen.

Outline

Suffix/Prefix Alignment Problem

Input: A set of r strings of total length n. Threshold t.

Output: For each string-pair, suffix/prefix matches of length $\geq t$.

A suffix/prefix match (*overlap*):

```
VÄLIMÄKI
    ||||
MÄKINEN
```

Motivation

Approximating the shortest common superstring.

Suffix/Prefix Alignment Problem

Input: A set of r strings of total length n. Threshold t.

Output: For each string-pair, suffix/prefix matches of length $\geq t$.

A suffix/prefix match (*overlap*):

```
VÄLIMÄKI
    ||||
   MÄKINEN
```

Motivation

Approximating the shortest common superstring.

Exact Overlaps

Longest overlaps in optimal time [Gusfield & Landau & Schieber, 1992]

- $O(n + output)$ time, $O(n)$ words.

Practical solution [Ohlebusch & Gog, 2010]

- $O(n + output)$ time, $8n$ bytes.

Finding *irreducible* overlaps [Simpson & Durpin, 2010]

- $O(n + output)$ time, $2nH_k + o(n) + r \log r$ bits.

Approximate Overlaps

Biological sequences have sequencing errors, SNPs...

First phase of *overlap-layout-consensus* assemby:

- ARACHNE [Batzoglou et al. 2002],
- Atlas [Havlak et al. 2004],
- Celera [Myers et al. 2000],
- Phrap [Green, 1994],
- UMD Overlapper [Roberts et al. 2004].

Their solution for approximate overlaps:

- Pick a *seed*, use exact matching and extend candidates,
- q-gram filters, expected $O(n^2/M)$ time [Myers, 2005] for a time–space tradeoff parameter M.

Approximate Overlaps

Biological sequences have sequencing errors, SNPs...

First phase of *overlap-layout-consensus* assemby:

- ARACHNE [Batzoglou et al. 2002],
- Atlas [Havlak et al. 2004],
- Celera [Myers et al. 2000],
- Phrap [Green, 1994],
- UMD Overlapper [Roberts et al. 2004].

Their solution for approximate overlaps:

- Pick a *seed*, use exact matching and extend candidates,
- q-gram filters, expected $O(n^2/M)$ time [Myers, 2005] for a time–space tradeoff parameter M.

Approximate Overlaps

We propose algorithms for

- k-errors: suffix/prefix edit distance $\leq k$,
- ϵ-errors: suffix/prefix edit distance $\leq \lceil \epsilon \ell \rceil$, where ℓ is the overlap length.

Workflow

1. preprocess $T_1, T_2, \ldots, T_r$ to build an *index*,
2. search each T_i separately.

Both steps can be parallelized.

Approximate Overlaps

We propose algorithms for

k-errors: suffix/prefix edit distance $\leq k$,

ϵ-errors: suffix/prefix edit distance $\leq \lceil \epsilon \ell \rceil$,
where ℓ is the overlap length.

Workflow

1. preprocess $T_1, T_2, \ldots, T_r$ to build an *index*,
2. search each T_i separately.

Both steps can be parallelized.

Virtues of the FM-Index

FM-index [Ferragina & Manzini, 2000] requires

$$nH_k(T) + o(n \log \sigma) \text{ bits}$$

and supports *backward searching*:

- Exact matching in $O(m)$ time [Ferragina et al. 2007],
- k-errors matching in $O(\sigma^k m^{k+1})$ time [Lam et al. 2008].

Where

- $H_k(T)$ is the k-th order entropy of T,
- m is pattern length,
- $\sigma = polylog(n)$ is alphabet size,
- locating occurrences takes extra time.

Overlaps with k-errors

Build the FM-index for:

$$T_1\$_1T_2\$_2\cdots T_r\$_r$$

and store permutation of $\$_i$'s in T^{bwt} in $r \log r$ bits.

- Prefix matches of P by searching $\$P$, in $O(m)$ time.

Prefix matches for all suffixes of T_i in $O(|T_i|)$ time

1. Iterate through T_i using backward search,
2. for each suffix, check if the SA interval can be extended with $\$$.

Overlaps with k-errors by plugging in [Lam et al. 2008].

Overlaps with k-errors

Build the FM-index for:

$$T_1\$_1T_2\$_2\cdots T_r\$_r$$

and store permutation of $\$_i$'s in T^{bwt} in $r \log r$ bits.

- Prefix matches of P by searching $\$P$, in $O(m)$ time.

Prefix matches for all suffixes of T_i in $O(|T_i|)$ time

1. Iterate through T_i using backward search,
2. for each suffix, check if the SA interval can be extended with $\$$.

Overlaps with k-errors by plugging in [Lam et al. 2008].

Overlaps with ϵ-errors

Simple pattern split strategy:

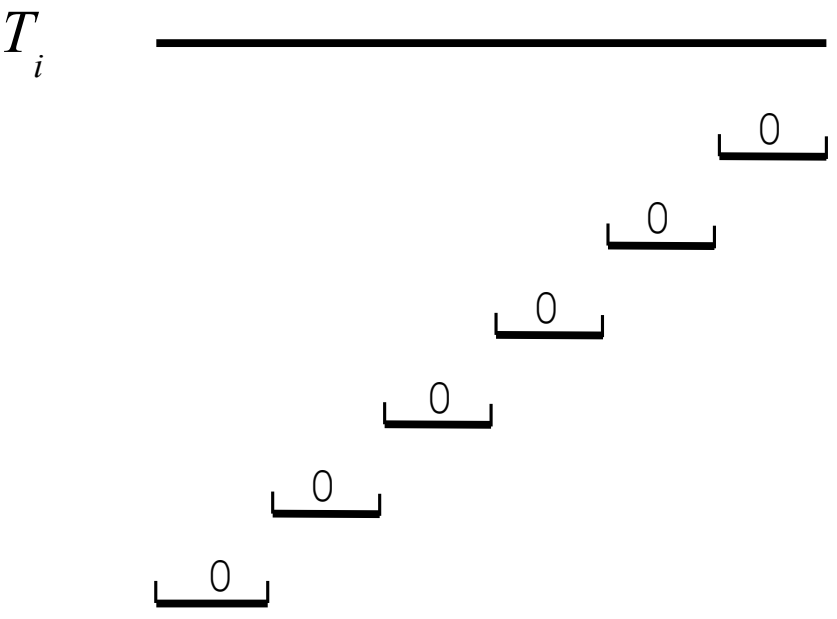
1. Split T_i into pieces of length

$$p = \min_{\ell=t}^{|T^i|} \left\lceil \frac{\ell}{\lceil \epsilon \ell \rceil + 1} \right\rceil$$

   Now at least one piece will match exactly.

2. Search each piece to find *candidate* overlaps,

3. verify candidate overlaps.



Overlaps with ϵ-errors

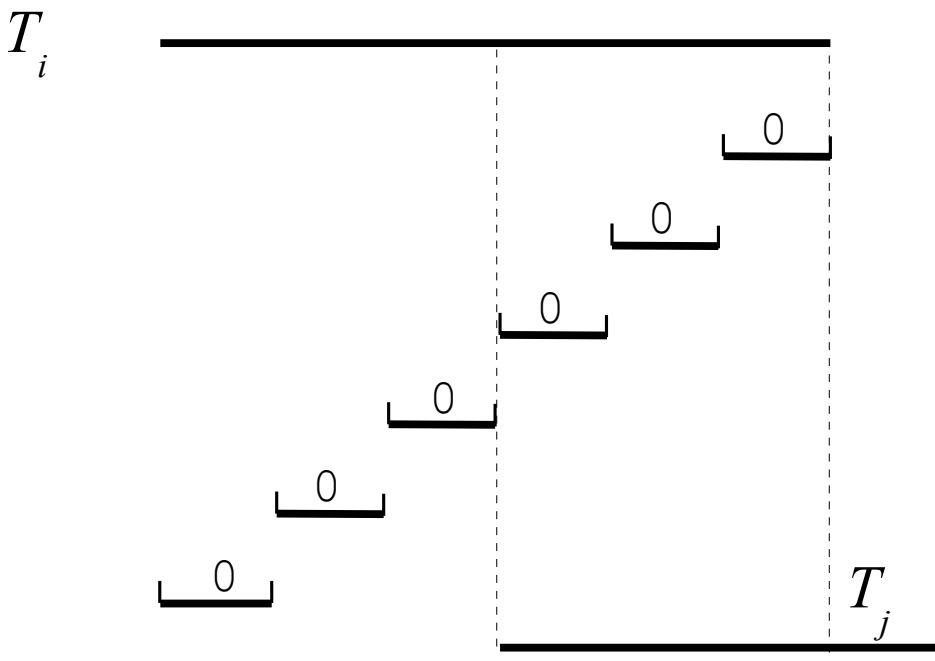
Simple pattern split strategy:

1. Split T_i into pieces of length

$$p = \min_{\ell=t}^{|T^i|} \left\lceil \frac{\ell}{\lceil \epsilon \ell \rceil + 1} \right\rceil$$

   Now at least one piece will match exactly.

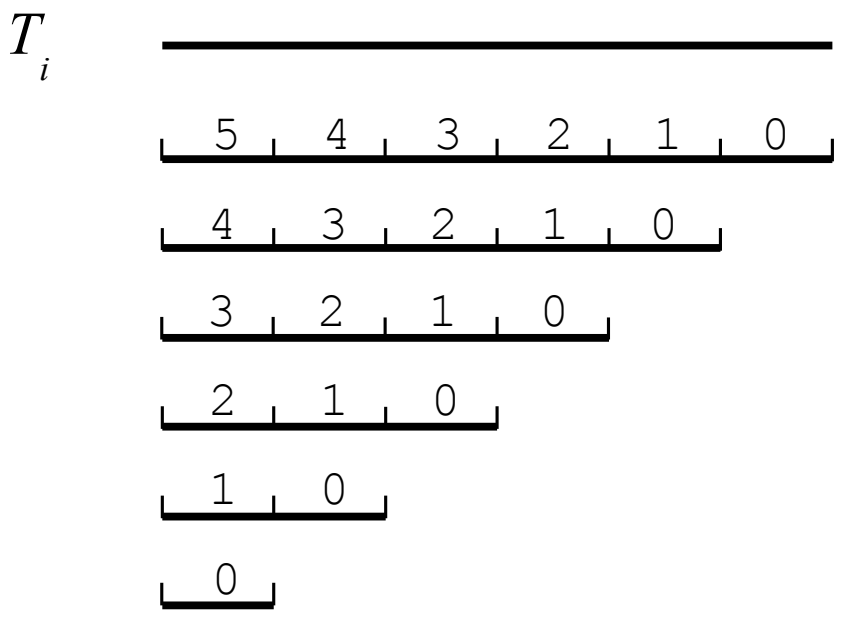
2. Search each piece to find *candidate* overlaps,
3. verify candidate overlaps.



Overlaps with ϵ-errors

Suffix filters [Kärkkäinen & Na, 2007]:

1. Split T_i into pieces of length p,
2. match with k-errors, and increase k at each boundary,
3. output candidate overlap only if search can be extended with \$.

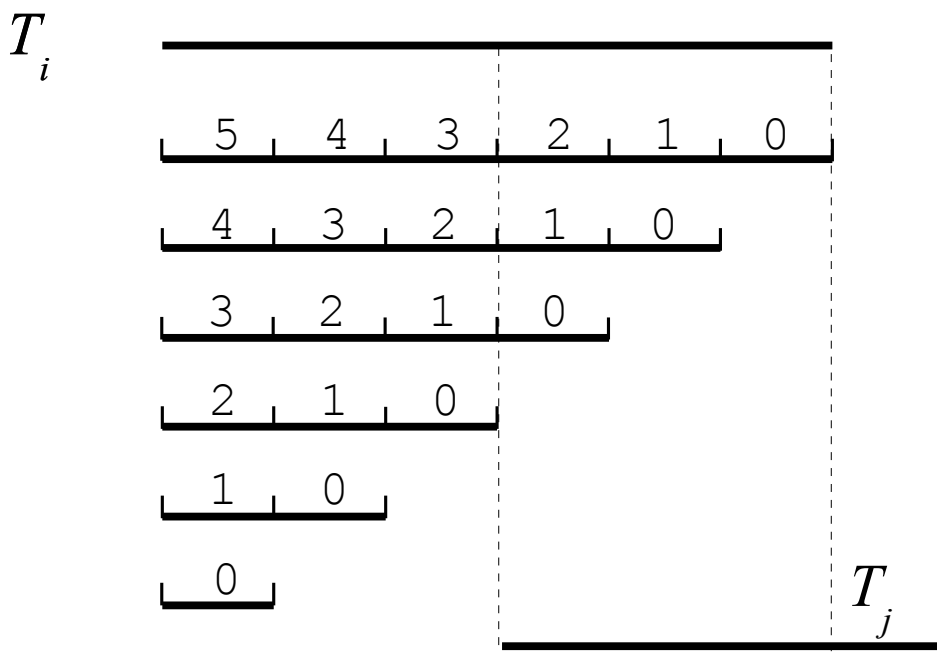


- *Stronger* filter criterion than pattern split [K & N, 2007],
- extending with \$ reduces the candidates even more.
- additional work needed to find all overlap lengths.

Overlaps with ϵ-errors

Suffix filters [Kärkkäinen & Na, 2007]:

1. Split T_i into pieces of length p,
2. match with k-errors, and increase k at each boundary,
3. output candidate overlap only if search can be extended with \$.



- *Stronger* filter criterion than pattern split [K & N, 2007],
- extending with \$ reduces the candidates even more.
- additional work needed to find all overlap lengths.

Melitaea Cinxia

De novo assembly of the *Glanville fritillary butterfly* genome:

- Metapopulation Research Group, University of Helsinki,
- Study of local populations 1993–,
- 4,000 habitats on 50 x 70 km area in the Åland Islands,
- 8 million 454 reads, average length around 350bp.

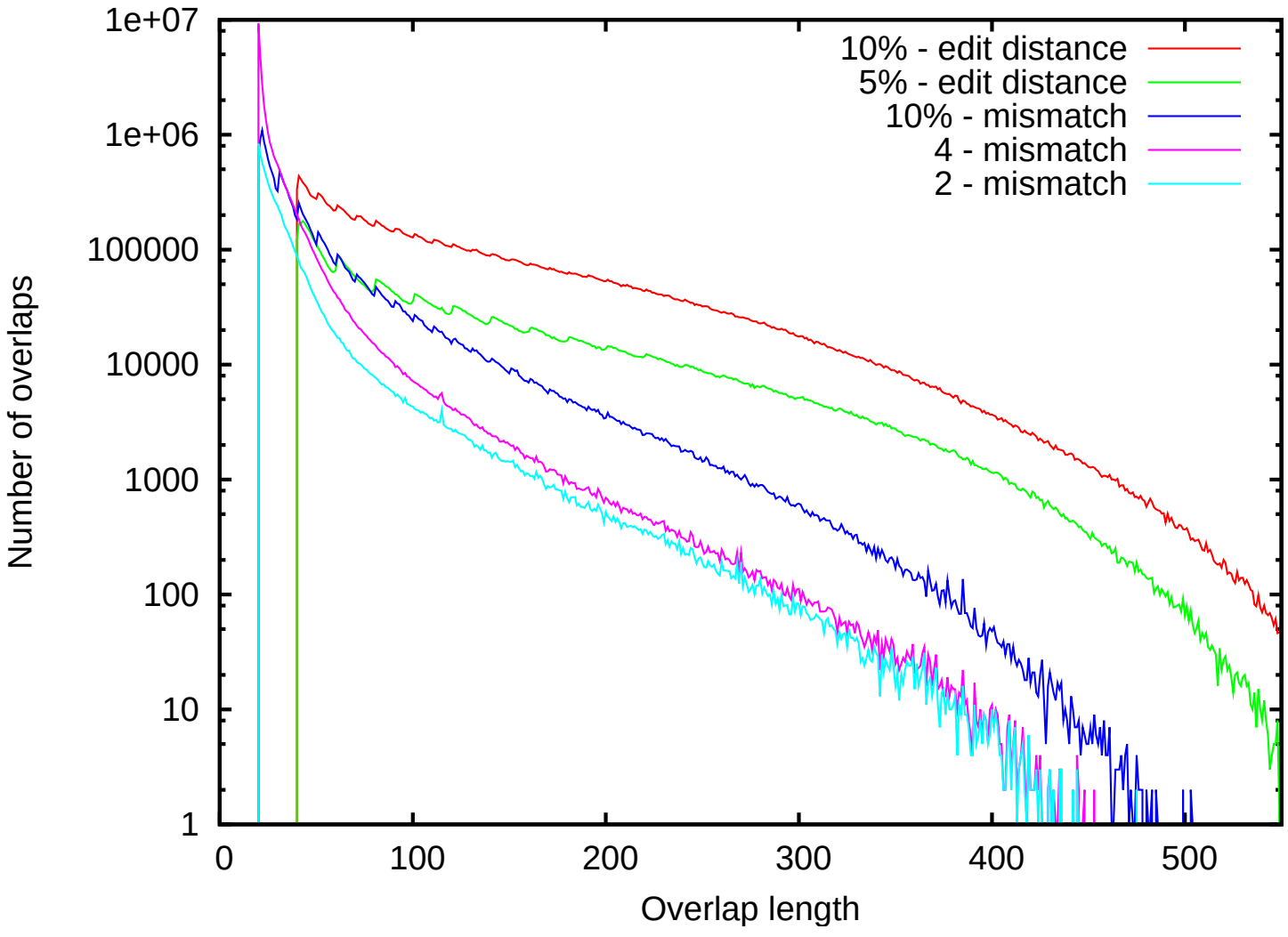
Results on 454 Reads

For a set of one million 454 reads:

Method	CPU time (h)	Max. ℓ	Avg. ℓ
2-mismatches	1.4	506	33.9
4-mismatches	76.9	506	27.4
2.5%-mismatches	8.0	506	74.8
5%-mismatches	14.7	524	76.7
2.5%-errors	20.0	561	116.1
5%-errors	49.7	1010	121.4

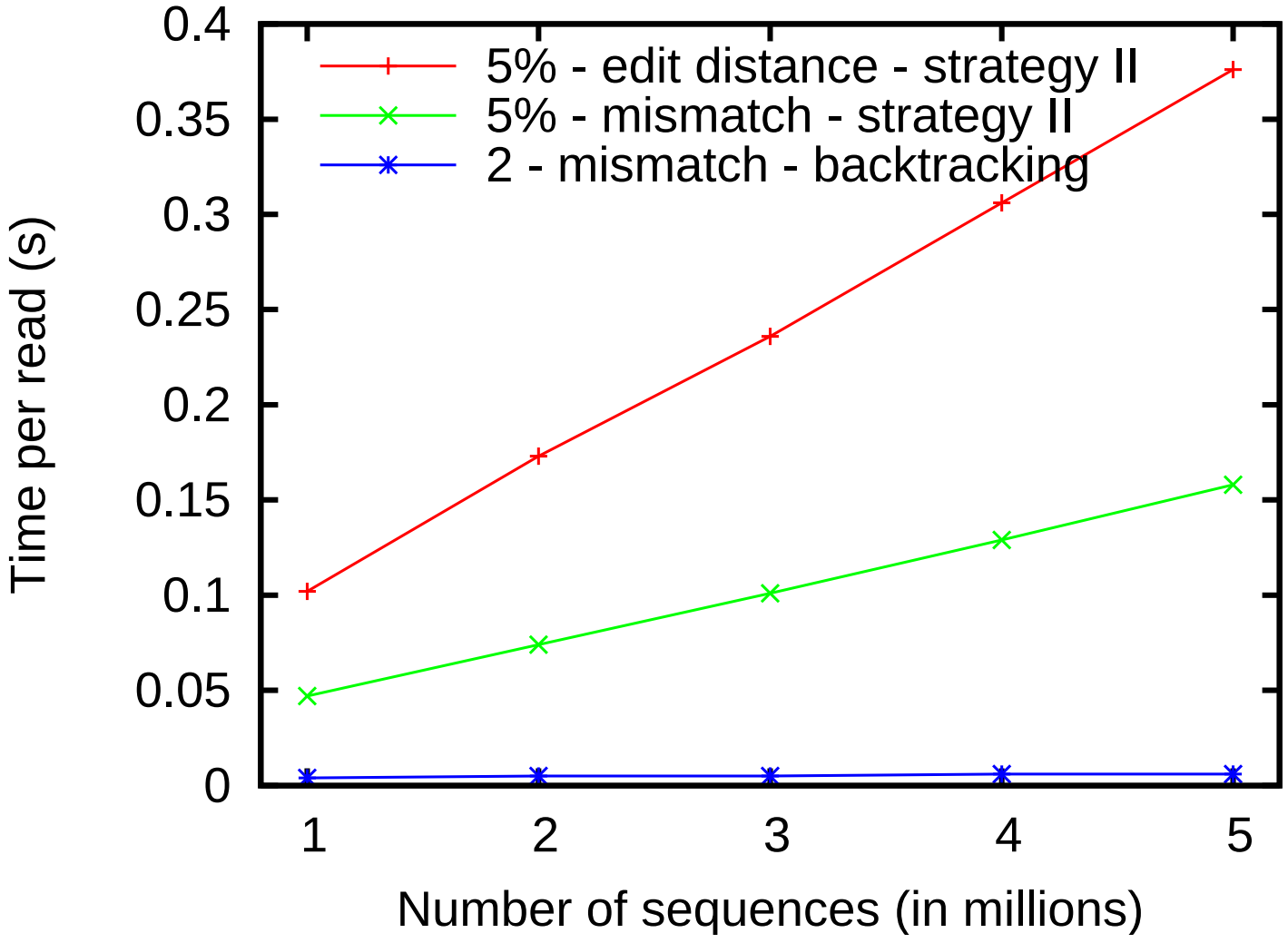
- $t = 20$ for k-mismatch, $t = 40$ for ϵ-errors.
- 454 pyrosequencing data contains mainly indel-errors.

Frequency of Overlap Lengths



- Light blue ≈ 1 CPU h, Green ≈ 48 CPU h.

When Number of Sequences Grows...



Future Plans

Adapt ideas from *approximate pattern matching*:

- *bidirectional* search [Lam et al. 2009], [Russo et al. 2009],
- $O(m + output + (\log n)^{k(k+1)} \log\log n)$ time,
  $O(n)$ space [Chan et al. 2006].

Comparison against

- q-gram filters [Myers, 2005],
- *de Bruijn graph* based assemblers.

Open problem: longest approximate overlaps.

Kiitos! Thank you!